SYSTEMS AND METHODS FOR DETERMINING AN UNKNOWN CHARACTERISTIC OF A SAMPLE

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is the U.S. National Phase of International Patent Application Serial No. PCT/US2014/059503 filed Oct. 7, 2014 which claims the benefit of priority under 35 U.S.C. § 119(e) of U.S. Provisional Application No. 62/055,446 filed Sep. 25, 2014 and U.S. Provisional Application No. 61/887,831 filed Oct. 7, 2013, the disclosures of which are hereby incorporated by reference in their entireties.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

This invention was made with government support under Grant No. 2011-DN-BX-K558 awarded by National Institute of Justice, Grant No. 2012-DN-BX-K050 awarded by the National Institute of Justice, and under Grant No. DBI-1126052 by the National Science Foundation. Accordingly, the Government has certain rights in this invention.

FIELD OF THE DISCLOSURE

The present disclosure relates to computerized analysis methods and systems to implement the computerized analysis methods. Specifically, the present disclosure relates to systems and methods for determining an unknown characteristic of a sample.

BACKGROUND OF THE INVENTION

Short Tandem Repeats, or STRs, are repetitive sequences 1-7 base pairs in length that are scattered throughout the human genome. One of the commonly used applications of STRs is in the field of human identification for forensic purposes. An STR DNA profile developed from a biological sample collected at a crime scene is compared with that of a person of interest or run against a database to check for a match. Biological evidence obtained at crime scenes is used to create a DNA profile and compared with the profile of a suspect to check whether a match occurs. In some instances, multiple people might have contributed to the evidence, giving rise to mixtures. The true number of contributors to a biological sample is never known with certainty. The DNA analyst is required to make assumptions about the number of contributors to the sample in order to reach a conclusion as to whether the suspect should be excluded or included as a potential contributor to the sample.

The Scientific Working Group on DNA Analysis Methods (SWGDAM) recommends that forensic reports include a statement as to the assumption made about the number, or the minimum number of contributors, to the sample being investigated. The number of contributors to a crime scene sample is generally unknown and must be estimated by the analyst based on the electropherogram obtained. The assumption on the number of contributors affects statistics used to assess the weight of DNA evidence, e.g., the Likelihood Ratio. Thus, it is useful to have a good estimate on the number of contributors to the sample.

Two commonly used methods to provide statistical weight for the inclusion of a person as a contributor are the Likelihood Ratio (LR) method and the Random Man Not Excluded (RMNE) method. Both of these methods require assumptions to be made concerning the number of contributors. Different assumptions lead to vastly different values for the LR method or different conclusions (i.e., inclusion or exclusion) in the case of the RMNE method. The most widely used method currently is Maximum Allele Count (MAC). This method seeks to identify the minimum number of individuals who could have contributed to a sample by counting the number of alleles observed at each locus, taking the maximum value over all the loci and dividing it by two.

Though methods to infer the number of contributors to a forensic sample exist, there are issues associated with all of them. Stochastic effects associated with DNA extraction, the PCR process and pipetting lead to non-detection of alleles (dropout). Further, allele sharing and PCR amplification artifacts like stutter occur frequently and make it difficult to interpret low-template, mixture profiles. These make it difficult to accurately estimate the number of contributors to a sample. The MAC method does not work well with complex mixtures because of sharing of alleles between the contributors. Guidelines have been established for estimating the number of contributors for high and low template samples using the total number of alleles observed. This method is prone to misclassification due to extensive allele sharing, dropout and stutter. Methods that do not rely only upon the number of alleles observed but also use the frequencies of the alleles in the signal have been created. For example, one method employing a Bayesian network has been developed and utilizes a probabilistic approach to infer the number of contributors to forensic samples. This method has been shown to work better than MAC with degraded DNA and with higher number of contributors. A Maximum Likelihood Estimator (MLE) method has also shown to give more accurate results than MAC with higher number of contributors and degraded DNA. A Probabilistic Mixture Model can infer the number of contributors to a sample based on the frequencies of the alleles observed.

SUMMARY OF THE INVENTION

A method and system is disclosed that takes a profile of an unknown sample as input, along with an amount of the sample, a set of calibration data, and a set of experimental conditions to determine an unknown characteristic of the unknown sample. The method and system then returns likelihoods for the number of contributors to the sample. This method and system uses quantitative data (e.g., peak heights in the signal) to estimate the number of contributors. In addition, it also uses the frequencies of the alleles observed. The method and system also incorporates stutter in its calculation. Probability of dropout is used in the calculation, as well as the various possible mixture ratios.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments will be described with reference to the following drawing figures, in which like numerals represent like items throughout the figures, and in which:

FIG. **1** is a flow chart that is useful for understanding method for determining an unknown characteristic of a sample;

FIG. **2** is a flow chart that is useful for understanding a method for generating calibration data;

FIG. **3** is a flow chart that is useful for understanding a method for analyzing a sample using generated calibration data; and